First passage times in homogeneous nucleation: Dependence on the total number of particles

Romain Yvinec, Samuel Bernard, Erwan Hingant, Laurent Pujo-Menjouet

First passage times in homogeneous nucleation: dependence on the total number of particles

Romain Yvinec[1], Samuel Bernard[2,3], Erwan Hingant[4], Laurent Pujo-Menjouet[2,3]
[1] *PRC INRA UMR85, CNRS UMR7247, Université François Rabelais de Tours, IFCE, F-37380 Nouzilly*
[2] *Université de Lyon, CNRS, Université Lyon 1, Institut Camille Jordan UMR5208, 69622 Villeurbanne, France*
[3] *INRIA Team Dracula, Inria Center Grenoble Rhne-Alpes, France and*
[4] *Departamento de Matemática, Universidad Federal de Campina Grande, PB, Brasil.*
(Dated: October 19, 2015)

Motivated by nucleation and molecular aggregation in physical, chemical and biological settings, we present an extension to a thorough analysis of the stochastic self-assembly of a fixed number of identical particles in a finite volume. We study the statistic of times it requires for maximal clusters to be completed, starting from a pure-monomeric particle configuration. For finite volume, we extend previous analytical approaches to the case of *arbitrary size-dependent* aggregation and fragmentation kinetic rates. For larger volume, we develop a scaling framework to study the behavior of the first assembly time as a function of the total quantity of particles.
We find that the mean time to first completion of a maximum-sized cluster may have surprisingly a very weak dependency on the total number of particles. We highlight how the higher statistic (variance, distribution) of the first passage time may still help to infer key parameters (such as the size of the maximum cluster) from data. And last but not least, we present a framework to quantify the formation of cluster of macroscopic size, whose formation is (asymptotically) very unlikely and occurs as a large deviation phenomenon from the mean-field limit. We argue that this framework is suitable to describe phase transition phenomena, as *inherent infrequent stochastic processes*, in contrast to classical nucleation theory.

PACS numbers: 02.50.Ga, 82.60.Nh, 87.10.Mn, 87.10.Rt

I. INTRODUCTION

The self-assembly of macromolecules and particles into cluster is a fundamental process in many physical, chemical an biological systems. Although particle nucleation and assembly have been studied for many decades[1,2], interest in this field has recently intensified due to engineering, biotechnological and imaging advances at the nanoscale level[3–5]. Applications range from material physics to cell physiology and virology (for a detailed list of examples see[6] and references therein). Many of these applications involve a fixed "maximum" cluster size – of tens to hundreds of units – at which the process is completed or beyond which the dynamics change[7,8]. One example include the rare self-assembly of mis-folded proteins into fibril aggregate in neurodegenerative diseases (Alzheimer, Parkinson, Prion...)[9,10]. Developing a stochastic self-assembly model focusing on the formation of a fixed "maximum" cluster size is thus important for our understanding of a large class of biological processes, and the quantification of the variability of the experimental data[11–15].

Theoretical models for self-assembly have typically described mean-field concentrations of clusters of all possible sizes using the well-studied mass-action, Becker-Döring equations[16–19]. While Master equations for the fully stochastic nucleation and growth problem have been derived, and initial analyses and simulations performed[20–24], there has been relatively less work on the stochastic self-assembly problem. It has been recently shown that in finite systems, where the maximum cluster size is capped, results from mass-action equations are inaccurate and that in this case a discrete stochastic treatment is necessary[6,25]. We consider here the Becker-Döring model (BD) defined by the following biochemical reactions

$$C_1 + C_i \underset{q_{i+1}}{\overset{p_i}{\rightleftharpoons}} C_{i+1}\,, \quad i \geq 1\,, \tag{1}$$

where C_i denotes the number (or concentration) of clusters of size i. Note that in the stochastic version (SBD), the state-space of the system is discrete and finite (see Fig. 1), given by all possible combinations that have a given fixed total number of particles (defined by M, given by the initial condition)

$$\mathcal{E} := \left\{ (C_i)_{i\geq 1} \subset \mathbb{N} \,:\, \sum_{i\geq 1} iC_i = M \right\}. \tag{2}$$

The configuration $(C_i(t))_{i\geq 1}$ evolve in continuous time by discrete jumps according to Markovian description of the reactions (1). In our previous examination of first assembly time in this model[6], we found that a striking finite-size effect arises in the limit of slow self-assembly. In particular, a *faster* detachment rate can lead to a *shorter* assembly time. This unexpected effect arise as the finite-size system may occupy some configurations that have been named "traps", where no free particle is available and the maximal-size cluster completion may occur only through first a detachment of a particle from a cluster. Discrepancies between the mean-field mass-action approach and the stochastic model were indeed

most apparent in the strong binding limit. In this paper, we will be interested in the generalization of earlier results[6] on the distribution of the first assembly times towards the completion of a full cluster, for arbitrary aggregation and fragmentation rates. Indeed, constant-size reaction rates was the main limitation of previous studies[6], as both physical and biological modeling require in many cases size-dependent attachment and detachment rates[13,26]. Moreover, we will focus here on the dependency of the assembly times on the total initial number of monomers M. We will show how statistics of the first assembly time as a function of the total number of monomers M may shed light on the biophysical properties of the appeared critical aggregates (size and size-dependence reaction rates). And we will highlight the discrepancies between the mean-field mass-action approach and the stochastic model *even in the presence of a high number of monomers M and its large limit.*

In the next section, we review the Becker-Döring mass-action equations for self-assembly and introduce the full stochastic problem. We derive the stochastic equations for the time-dependent cluster numbers, and introduce assembly times as first passage time problems. In Section III, we explore two simplified models for which the first assembly time can be solved analytically, and derive asymptotic expressions for the first assembly time in both the large number of monomer limit and large cluster size limit. Results from kinetic Monte-Carlo simulations (or Gillespie's algorithm) are presented in Section IV and compared with our analytical estimates. Finally, we discuss the implications of our results and propose further extensions in the Summary and Conclusions.

II. STOCHASTIC BECKER-DÖRING MODEL, FIRST ASSEMBLY TIMES DEFINITIONS

The classic deterministic mass-action description for spontaneous, homogeneous self-assembly is the Becker-Döring model[1] (BD), where the concentrations $c_k(t)$ of clusters of size k obey an infinite (or truncated up to $k = N$) system of ordinary differential equation, given, for all $t \geq 0$, by

$$\left\{ \begin{array}{rcl} \dfrac{d}{dt}c_1(t) & = & -2j_1(t) - \displaystyle\sum_{k\geq 2} j_k(t)\,, \\ \dfrac{d}{dt}c_k(t) & = & j_{k-1}(t) - j_k(t)\,, \quad k \geq 2\,, \end{array} \right. \tag{3}$$

with

$$j_k(t) = p_k c_1(t) c_k(t) - q_{k+1} c_{k+1}(t)\,, \quad k \geq 1\,, \tag{4}$$

and initial condition $c_1(0) = M$ and $c_k(0) = 0$ for all $k \geq 2$. The rates p_k and q_k are respectively the monomer attachment and detachment rates to and from a cluster of size k. These rates are limited to sub-linear function of k, with bounded increments, in order to fulfill the standard well-posedness criteria[27,28]. It has been previously shown that such equations provide a poor approximation of the expected number of clusters when the total mass M and the maximum cluster size N are comparable in magnitude[25]. Furthermore, such representations do not capture the randomness of the binding/unbinding events and cannot describe the heterogeneity of cluster size distribution and time-dependent property such as first assembly times. *A stochastic treatment is thus necessary and is the subject of the remainder of this paper.*

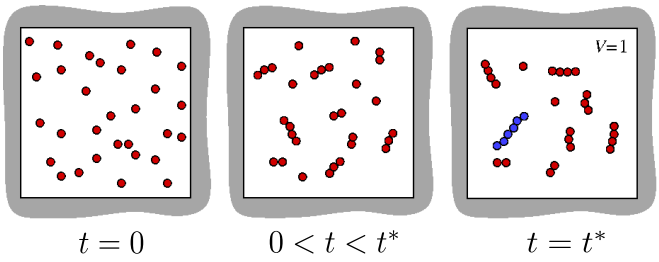

FIG. 1: Homogeneous self-assembly and growth in a closed unit volume initiated with $M = 30$ free monomers. At a specific intermediate time $0 < t < t^*$ in this depicted realization, there are six free monomers, four dimers, four trimers, and one cluster of size four. For each realization of this process, there will be a specific time t^* at which a maximum cluster of size $N = 6$ in this example is first formed (blue cluster). This time t^* is a realization of the First Assembly Time (FAT, see definition in (7))

Using a Markovian approach, we have previously derived[6] the Forward Master equation to describe the probability that the system is in any given admissible configuration (see Eq. (2)) at times $t \geq 0$. An equivalent formulation of this model is given by stochastic equations, driven by Poisson processes (as a continuous-time Markov Chain). This approach is natural to perform numerical simulations of sample path, and is more efficient to compute numerically first assembly times than the Master Equation approach. Moreover, this approach leads to natural comparison with deterministic systems when the total mass M is large. Denoting by Y_k^+ (resp. Y_k^-) the standard Poisson process associated to the forward aggregation (resp. fragmentation) reaction of clusters of size k, the stochastic Becker-Döring (SBD) equations for the time evolution of the number $C_k(t)$ of cluster of size k are given for $t \geq 0$ by (starting from a *pure monomeric initial condition*)

$$\left\{ \begin{array}{rcl} C_1(t) & = & M - 2J_1(t) - \displaystyle\sum_{k\geq 2} J_k(t)\,, \\ C_k(t) & = & J_{k-1}(t) - J_k(t)\,, \quad k \geq 2\,, \end{array} \right. \tag{5}$$

with

$$\begin{array}{rcl} J_k(t) & = & Y_k^+\Big(\displaystyle\int_0^t p_k C_1(s)(C_k(s) - \delta_k^1) ds \Big) \\ & & -Y_{k+1}^-\Big(\displaystyle\int_0^t q_{k+1} C_{k+1}(s) ds \Big)\,, \quad k \geq 1\,. \end{array} \tag{6}$$

where $\delta_k^1 = 1$ if $k = 1$ and $\delta_k^1 = 0$ if $k > 1$.
The first assembly time (FAT) for the stochastic discrete Becker-Döring is defined as a first passage time problem[29]

$$T_{1,0}^{N,M} := \inf\{t \geq 0 : C_N(t) = 1\}\,. \tag{7}$$

Hence the FAT is the first time to obtain a cluster of size N, starting with a pure single particle initial state, with M particle (see Fig. 1 for an example). To link with macroscopic definition of the nucleation time, we will also consider the generalized first assembly time (GFAT) problems

$$T_{\rho,h}^{N,M} := \inf\{t \geq 0 : C_N(t) \geq \rho M^h\}\,, \tag{8}$$

for given positive constant ρ and $0 \leq h \leq 1$. Here, we want to analyze the behavior of $T_{\rho,h}^{N,M}$ when $M \to \infty$, for fixed N, and when both $M, N \to \infty$. This behavior will depend on scaling on the physical rates p, q, which may naturally depend on the total mass (or volume) of the system[30]. One way of computing the distribution of first assembly times is to consider the Backward Kolmogorov equation (BKE) describing the evolution of the configuration probabilities as a function of local changes from the initial configuration. The BKE Approach was taken in[6]. It has the advantage to yields exact results for the full distribution of FAT, but it is strictly limited by the size of the system of equations, that grows exponentially with M. In this paper, we rely on exact calculation of simplified reduced models, limit theorems from Eq. (5) for large M and N, and extensive numerical simulations of these equations.

III. RESULTS AND ANALYSIS

Although the state-space (2) of the SBD model (1) is finite, the first passage problem defined by Eq. (8) is in general a very difficult problem: see preliminary studies in[6,20,21,23]. There are two distinct simplifications that allow the problem to be analytically tractable. We present them here briefly in sections **A** and **B** (and generalize results from[6]). Then we present two asymptotic results for large volume, $M \to \infty$, with either finite or infinite nucleus size in sections **C** and **D**, respectively. The strategy will be based on re-scaling procedure of the stochastic Eq. (5). Numerical illustrations and results are postponed to the next section.

A. Constant Monomer formulation

The SBD model defined by Eq. (5) has the constant mass property

$$\sum_{k\geq 1} kC_k(t) \equiv \sum_{k\geq 1} kC_k(0) = M\,, \quad t \geq 0\,.$$

In the original formulation of the BD model (sometimes used in the deterministic context[17,27]), the total mass of the system is not preserved, but rather the quantity of free particles (think *e.g.* of a source/sink that will keep instantaneously constant the quantity of available free particles). We will refer to this formulation as the constant monomer stochastic Becker-Döring model (CMSBD). We can represent it by the following reactions

$$\begin{cases} \emptyset \quad \underset{q_2}{\overset{p_1 M(M-1)}{\rightleftharpoons}} \quad C_2\,, \\ C_k \quad \underset{q_{k+1}}{\overset{p_k M}{\rightleftharpoons}} \quad C_{k+1}\,, \quad i \geq 2\,. \end{cases} \tag{9}$$

In the above formulation, $C_1(t) \equiv M$ is now a constant over time. Note that we expect such model to be close to the original SBD (for small times, up to the FAT) in the limit of large number of particle M. The main advantage of the constant monomer formulation is to be linear and hence analytically solvable. Indeed, it is known that for linear population model[31], the number of individuals in each subclass of the population (starting with no individuals at time 0), namely here $C_2(t), ...C_N(t)...$, are independent Poisson random variable. Moreover, for this model (9), the mean $c_2(t), ...c_N(t)...$ are solution of the linear equation, given for all $t \geq 0$, by

$$\frac{d}{dt}c_k(t) = j_{k-1}(t) - j_k(t)\,, \quad \forall k \geq 2\,, \tag{10}$$

with

$$\begin{cases} j_1(t) &= p_1 M(M-1) - q_2 c_2(t)\,, \\ j_k(t) &= p_k M c_k(t) - q_{k+1} c_{k+1}(t) \quad \forall k \geq 2\,, \end{cases} \tag{11}$$

and initial condition $c_k(0) = 0$ for all $k \geq 2$. Note that the last set of Eq. (10)-(11) is very close to the deterministic Becker-Döring model (3)-(4) taking $c_1 \equiv M$. To calculate the FAT $T_{1,0}^{N,M}$ we use the survival function

$$\begin{aligned} S_{1,0}^{N,M}(t) &:= \mathbb{P}\big\{T_{1,0}^{N,M} > t\big\} \\ &= \mathbb{P}\big\{C_N(s) = 0, s \leq t \mid C_k(0) = M\delta_{k=1}\big\}\,. \end{aligned}$$

Then, using an absorbing boundary condition at $k = N$ ($q_N = p_N = 0$) together with the initial condition entail that $C_N(t) = 0$ for some $t \geq 0$ if and only if $C_N(s) = 0$ for all $s \leq t$, so that

$$S_{1,0}^{N,M}(t) = \mathbb{P}\big\{C_N(t) = 0 \mid C_k(0) = M\delta_{k=1}\big\}\,.$$

Finally, since $C_N(t)$ is Poisson distributed (linear system) with mean $c_N(t)$, we have

$$S_{1,0}^{N,M}(t) = e^{-c_N(t)}\,. \tag{12}$$

Equations (10)-(11) with the absorbing boundary at $k = N$ can rewritten as a linear system

$$\begin{cases} \dot{\mathbf{c}} &= \mathbf{A}\mathbf{c} + \mathbf{B}\,, \\ \dot{c}_N(t) &= p_{N-1} M c_{N-1}(t)\,, \end{cases} \tag{13}$$

where $\mathbf{c} = (c_2, c_3, \cdots, c_{n-1})^T$, $\mathbf{B} = (p_1 M(M-1), 0, \cdots, 0)^T$ and $\mathbf{A}$ is a tridiagonal matrix with elements

$$\left\{ \begin{array}{rcl} a_{k,k} & = & -q_{k+1} - p_{k+1} M \,, \\ a_{k+1,k} & = & p_{k+1} M \,, \\ a_{k,k+1} & = & q_{k+1} \,. \end{array} \right.$$

The study of the linear system (13) has been performed both for the infinite dimensional case[32] and for the truncated case[33]. In particular, it is shown that a similarity transformation

$$\mathbf{A} = \mathbf{D}\mathbf{S}\mathbf{D}^{-1}$$

with $D = diag(\sqrt{\tilde{Q}_k M^k})$ and $\tilde{Q}_k = \prod_{j=2}^{k} \frac{p_{j-1}}{q_j}$ leads to a matrix $\mathbf{S}$ real symmetric tri-diagonal with non-zero elements on the sub and super-diagonal. Hence, classical linear algebra results shows that the eigenvalues of $\mathbf{A}$ are real and distinct. Then, a general form of $c_N(t)$ is given by

$$c_N(t) = p_{N-1} M \Big[\sum_{k=1}^{N-2} \alpha_k V_{N-2}^{(k)} \frac{e^{\lambda_k t} - 1}{\lambda_k} - (\mathbf{A}^{-1}\mathbf{B})_{N-2} t \Big].$$

where λ_k, V^k denotes the eigenelements of $\mathbf{A}$ and α_k are determined by initial conditions. Analytical solutions are available[12] for constant coefficient only (the matrix $\mathbf{A}$ is in such case a Toeplitz matrix, with constant diagonal values, see Annex A.1). However, asymptotic expression are valid in the general cases. In particular, we have for small times $c_2(t) = p_1 M(M-1)t + o(t)$ and

$$\dot{c}_k = p_{k-1} M c_{k-1} + o(t) \,,$$

so that, for $t \ll 1$,

$$c_N(t) \approx_{t \ll 1} M^N \prod_{k=1}^{N-1} p_k \frac{t^{N-1}}{(N-1)!} \,,$$

and Eq. (12) is thus the survival function of a Weibull distribution, of shape parameter $k = N-1$ and scale parameter $\lambda = ((N-1)!/(M^N \prod_{k=1}^{N-1} p_k))^{1/(N-1)}$. Hence, we get

$$\langle T_{1,0}^{N,M} \rangle \approx_{M \to \infty} \frac{\Gamma(1 + 1/(N-1))}{\Big(\prod_{k=1}^{N-1} p_k \Big)^{1/(N-1)}} \frac{((N-1)!)^{1/(N-1)}}{M^{1+1/(N-1)}} \,. \tag{14}$$

Variance formula for the Weibull distribution yields the asymptotic coefficient of variation (standard deviation over the mean)

$$cv_{T_{1,0}^{N,M}} \approx_{M \to \infty} \sqrt{2(N-1) \frac{\Gamma(2/(N-1)}{\Gamma(1/(N-1))^2} - 1} \,. \tag{15}$$

Note in particular that the coefficient of variation do not vanish in large population, and that it is independent of the particular shape of the aggregation rate and depends only on the size of the maximal cluster N.

For the generalized first assembly time GFAT, similar time scale asymptotic on the Eq. (13) on the mean gives the following expression

$$\langle T_{\rho,h}^{N,M} \rangle \approx_{M \to \infty} \frac{C(p,N)}{M} \frac{1}{M^{(1-h)/(N-1)}} \,, \tag{16}$$

where $C(p,N)$ is a constant that depends only on N and the aggregation rates $p_k, k \leq N$ (that can be made explicit if the full solution of Eq. (13) is known). Those asymptotic expressions are illustrated in Annex (Fig. A.1) where a perfect match is observed with numerical simulations.

B. Single cluster model

Another simplified model that can be analytically solved for the FAT problem is given by the assumption that a single cluster can be formed at a time[6,34]. We expect such model to be close to the original model when the fragmentation dominates, so that formation of many (large) cluster is unlikely. In such model, called the single-cluster stochastic Becker-Döring (SCSBD) model we may represent only the size of the single cluster, so that the state space is now one dimensional, being simply

$$\mathcal{E}_1 := [1, \ldots, N] \,,$$

and the possible reactions are given by (k denotes the size of the single cluster)

$$\left\{ \begin{array}{rcl} k = 1 & \underset{q_2}{\overset{p_1 M(M-1)}{\rightleftharpoons}} & k = 2 \,, \\ k & \underset{q_{k+1}}{\overset{p_k (M-k)}{\rightleftharpoons}} & k+1 \,, \quad k \geq 2 \,. \end{array} \right. \tag{17}$$

In such a scenario, exact solution and classical First Passage Theory[30] gives (it is a one-dimensional discrete random walk)

$$\langle T_{1,0}^{N,M} \rangle = \sum_{i=1}^{N-1} \sum_{j=1}^{i} \frac{\prod_{k=j+1}^{i} q_k}{\prod_{k=j}^{i} p_k} \frac{1}{M^{\delta_j^1} \prod_{k=j}^{i} (M-k)} \,. \tag{18}$$

In addition, general formula for variance and cumulative distribution function are available[35]. Those theoretical expressions are illustrated in Annex (Fig. A.2) where a perfect match is observed with numerical simulations.

Asymptotic expressions of the mean assembly time is straightforwardly deduced from Eq. (18). For instance, assume $q_k = \frac{\overline{q}_k}{\varepsilon}$ and that $\varepsilon \to 0$, the leading order of the mean assembly time is

$$\langle T_{1,0}^{N,M} \rangle \approx_{\varepsilon \to 0} \frac{1}{\varepsilon^{N-2}} \frac{\prod_{k=2}^{N-1} \overline{q}_k}{\prod_{k=1}^{N-1} p_k \prod_{k=0}^{N-1} (M-k)} \,.$$

Also, one can show that in the asymptotic $\varepsilon \to 0$, for large fragmentation rate, the FAT $T_{1,0}^{N,M}$ is an exponential distribution[6].

Finally, for large N and M, we can rescale the sum in Eq. (18) to obtain a suitable expression for the mean FAT when $N \to \infty$. Assume that the aggregation rates scale with M so that $p_1 = \overline{p}_1/M^2$, $p_k = \overline{p}_k/M$, $k \geq 2$. Then, let us introduce the rescaled size variable $x = k/N$, and define the rescaled kinetic rate

$$\begin{aligned} \overline{p}(x) &= \sum_{k\geq 2} \overline{p}_k \mathbf{1}_{[k/N,(k+1)/N)}(x)\,, \\ q(x) &= \sum_{k\geq 2} q_k \mathbf{1}_{[k/N,(k+1)/N)}(x)\,, \end{aligned}$$

we have, for $N = \sqrt{M} \to \infty$ (see details in Annex, section A.2.2),

$$\langle T_{1,0}^{\sqrt{M},M} \rangle \approx_{M\to\infty} M \int_0^1 \int_0^y \frac{e^{(y^2-z^2)/2}}{q(y)} \cdot \exp\Big[\sqrt{M}\int_z^y \ln\left(\frac{q(x)}{\overline{p}(x)}\right) dx\Big] dydz\,. \quad (19)$$

In particular, when $q(x) > \overline{p}(x)$ on an interval of positive measure on $[0,1]$, the last expression (19) implies that the mean FAT to reach the *macroscopic* size $x = 1$ ($k = N = \sqrt{M}$) is exponentially large as $M \to \infty$. As an example, suppose that $q > \overline{p}$ are size-independent. Then, Eq. (19) becomes

$$\langle T_{1,0}^{\sqrt{M},M} \rangle \approx_{M\to\infty} \frac{M}{q} \int_0^1 \int_0^y e^{(y^2-z^2)/2} \left(\frac{q}{\overline{p}}\right)^{\sqrt{M}(y-z)} dydz\,.$$

Those theoretical expressions are illustrated in Annex (Fig. A.3 and A.4). Note that a different approach is to link the one-dimensional discrete random walk (17) with a one-dimensional stochastic differential equation, and to use Large Deviation Theory to derive asymptotic FAT[34] (Annex, section A.2.3). This scaling approach and the link with a continuous size model when $N \to \infty$ will be taken on the full SBD model in section **D**.

C. Large M, finite N

In this section, we investigate the behavior of the SBD and its FAT when the total number of particles M tends to infinity, while the size N of the maximal cluster to reach stay finite. We distinguish two scenario, which yields distinct results. In the first one, the aggregation and fragmentation rate p_k, q_k are taken independent of M. As the aggregation propensities increase with M, it is expected that the FAT decrease to 0 as $M \to \infty$. The objective is to find valid asymptotic expression, and its dependence with respect to other parameters, like the maximal cluster size N for instance. In the second case, the aggregation rate is scaled with the total number of particles, with $p_k = \frac{\overline{p}_k}{M}$. This scaling is motivated by classical system size expansion of chemical reaction networks[30]. As the total number of particles increases, the volume also increases so that the overall reaction propensities of the aggregation reactions stay constant. In such case, one expect to recover the deterministic first passage time of the classical deterministic BD model.

Let us now introduce our general rescaling strategy. The number of cluster of size k, given by C_k, are rescaled into

$$D_k^M(t) = \frac{C_k(t/M^\gamma)}{M}$$

with γ a scaling coefficient to be chosen latter. Then, from Eq. (5)-(6), we obtain, for any $t \geq 0$,

$$\begin{cases} D_1^M(t) &= 1 - 2J_1^M(t) - \sum_{k\geq 2} J_k^M(t)\,, \\ D_k^M(t) &= J_{k-1}^M(t) - J_k^M(t)\,, \quad k \geq 2\,, \end{cases} \quad (20)$$

with

$$J_k^M(t) = \frac{1}{M} Y_k^+\Big(\int_0^t M^{2-\gamma} p_k D_1^M(s)(D_k^M(s) - M^{-1}\delta_k^1) ds\Big) - \frac{1}{M} Y_{k+1}^-\Big(\int_0^t M^{1-\gamma} q_{k+1} D_{k+1}^M(s) ds\Big)\,, \quad k \geq 2\,. \quad (21)$$

We recall a standard result of convergence of Poisson Processes (law of large numbers[36]), that

$$\frac{1}{n} Y(nt) \rightharpoonup_{n\to\infty} t\,,$$

where Y is a standard Poisson Process.

1. No scaling of the aggregation rate

Using $\gamma = 1$, and the standard law of large numbers applied to the Eq. (20)-(21), we can show[37] (see Annex, section 3) that the sequence of stochastic processes $(D_k^M(t))$ converges, as $M \to \infty$, in a rigorous sense (in the trajectory space) to the solution of the irreversible aggregation deterministic model (BD with $q_k = 0$), given, for all $t \geq 0$, by

$$\begin{cases} \frac{d}{dt} d_1 &= -2j_1(t) - \sum_{k\geq 2} j_k(t)\,, \\ \frac{d}{dt} d_k &= j_{k-1}(t) - j_k(t)\,, \qquad \forall k \geq 2\,, \end{cases} \quad (22)$$

with

$$j_k(t) = p_k d_1 d_k(t)\,, \forall k \geq 1\,, \quad (23)$$

and initial condition $d_1(0) = 1$ and $d_k(0) = 0$, for all $k \geq 2$. Intuitively, in the rescaled variable D_k^M, the aggregation process is much more favorable compared to the fragmentation because the number of free particles is very large. By definition of the GFAT Eq. (8), with $h = 1$,

$$MT_{\rho,1}^{N,M} = \inf\{t \geq 0 : D_N^M(t) \geq \rho\}\,.$$

Then, using the convergence of $(D_k^M(t))$, we obtain the following asymptotic behavior of the GFAT for $h = 1$,

$$\lim_{M\to\infty} MT_{\rho,1}^{N,M} = \inf\{t \geq 0 : d_N(t) \geq \rho\}\,. \qquad (24)$$

The latter quantity is deterministic, and may be finite or infinite, according to the respective value of p_k, N and ρ.

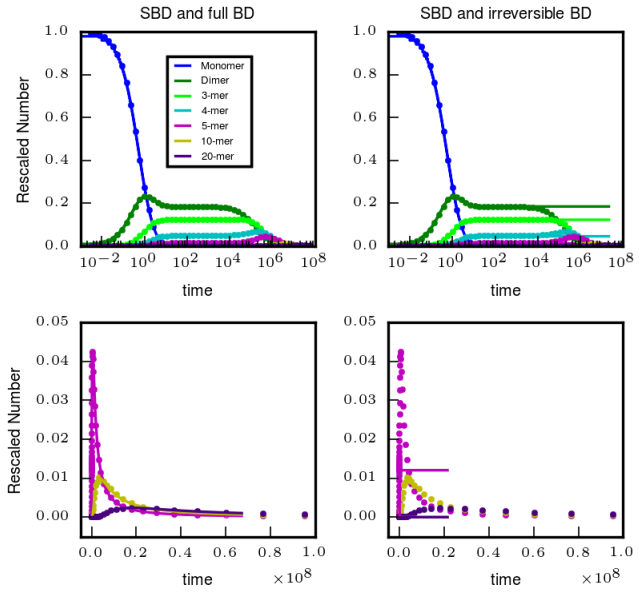

FIG. 2: SBD and BD Trajectories. For the SBD, we simulate the rescaled Eq. (20)-(21), with $M = 10^{5.5}$, and kinetic rates are $p_k \equiv 1$ for all $k \geq 1$ and $q_k \equiv 1$ for all $k \geq 2$. For the BD model, we simulate on the left columns the full BD Eq. (22)-(26) and on the right columns the irreversible BD Eq. (22)-(23). The rescaled SBD trajectories are plotted with filled circles, together with the corresponding BD trajectories in plain lines, for the monomer and i-cluster, $i = 2, 3, 4, 5, 10, 20$, according to the legend. The lower panel correspond to the same numerical simulation of the upper panel, with a zoom on the y-axis to improve the visaulization of the $i-cluster$ for $i = 5, 10, 20$. It is immediate to see that the full BD Eq. (22)-(26) agrees perfectly with the rescaled SBD Eq. (20)-(21) for all time, while the irreversible BD Eq. (22)-(23) matches only up to a time scale of order M.

The limit model (22)-(23) do not capture the FAT and the GFAT $T_{\rho,h}^{N,M}$ for $h < 1$ (such event is reached for time $t = 0^+$). However, as the initial number of monomers is large, we can use an intermediate approximation of the dynamic of the stochastic model, as a hybrid deterministic/stochastic model. First, note that pure coagulation BD model (22)-(23) have been extensively studied[38], where exact time-dependent solution for $p_k = pk$ are given, and time asymptotic behavior are given for power law coefficient $p_k = pk^\lambda$, $0 \leq \lambda \leq 1$. We restrict the following discussion to the constant rate case, $\lambda = 0$, for simplicity (results are analogous in the power law case). In such case, the stationary state of the pure coagulation BD model[17,38] (22)-(23) is $d_1^* = 0$ and

$$d_k^* = \frac{k-1}{ek!}\,, \quad k \geq 2\,. \qquad (25)$$

Although the rescaled threshold ρM^{h-1} will be reached by d_N (and hence by D_N^M) for any ρ and $h < 1$ for large enough M (as $d_N^* > 0$), one can already see that for "intermediate" M, we may have $Md_N^* \ll 1$, so that the threshold may not have been reached while the free particle have vanished ($d_1^* = 0$). In such case, it is necessary to take into account the small but crucial contribution of the aggregate shortening. For that, let us consider as a further approximation of Eq. (20)-(21) the following deterministic model (still with constant rate coefficients, to simplify the following), given, for all $t \geq 0$, by Eq. (22) and flux definition

$$j_k(t) = pd_1(t)d_k(t) - \frac{1}{M}qd_{k+1}(t)\,, \quad k \geq 1\,, \qquad (26)$$

where $1/M$ is seen as a small parameter. To obtain results for the quantity $T_{\rho,h}^{N,M}$ for any $h < 1$, we need to study the time-dependent properties of the favorable aggregation limit $M \to \infty$ of the deterministic BD model Eq. (22)-(26). The following discussion is illustrated with Fig. 2 (see also in Annex, Fig. A.7, A.8). For constant rate p, q, it is known[17] that under favorable aggregation limit $q/M \ll p$, the deterministic BD model Eq. (22)-(26) exhibits the following successive periods:

- Firstly, the model behaves as the irreversible aggregation BD model, Eq. (22)-(23), during a time-scale of order $e\log(M)$, until monomer concentration $d_1(t)$ becomes small;
- Secondly, when the monomer concentration d_1 is of order $1/M$, there is a metastable period in which each concentration species of size $k \geq 2$ are nearly constant, equal to d_k^*, the equilibrium state Eq. (25) of the irreversible aggregation BD model, Eq. (22)-(23). The concentration $d_k(t)$ stays roughly constant to the values d_k^*, distinct from the steady-state values of the full BD Eq. (22)-(26), until the next time scale;
- Thirdly, at a time scale of order M (which is the time scale of aggregate shortening), larger aggregates are created within a process akin to diffusion in the size k-space (slow redistribution of aggregate sizes);
- Finally, every concentration species d_k reaches the classical steady-state value of the full BD Eq. (22)-(26) within a time scale of order M^2. Steady-state values $\overline{d}_k$ are given by

$$\overline{d}_k = \left(\frac{pM}{q}\right)^{k-1} \overline{d}_1^k\,, \quad k \geq 2\,,$$

where $\overline{d}_1$ is determined by the mass conservation property. Such values can be approximated by $\overline{d}_1 \approx 1/M$ and $\overline{d}_k \approx 1/M(1 - 1/\sqrt{M})^{k-1}$.

To approximate the GFAT $T_{\rho,h}^{N,M}$, we need to know in which of these periods the event $\{C_N(t) \geq \rho M^h\}$ is

reached. This mostly depends on the critical size N of the nucleus as follows. If M is large enough, then the metastable state is large enough, *i.e.* $c_N^* = Md_N^* > \rho M^h$, and the cluster number $C_N(t)$ will reach the threshold during the pure-aggregation time-scale ($\log(M)/M$ in the original time scale), and the GFAT $T_{\rho,h}^{N,M}$ is found (see numerical section) to behave as the linear CMSBD model (9) with $C_1 = M$.

In the opposite case, for intermediate M and large enough N such that $c_N^* \ll \rho M^h$, we expect $C_N(t)$ to reach the threshold after the metastable period (of order 1 in the original time scale). As the initial pure-aggregation phase is short ($\log(M)/M$), we can neglect it, and use the metastable values c_k^* as initial values for a linear CMSBD model (9) where the monomer number is now equal to $C_1 \equiv c_1^*$ (see Annex, section 3.1 and Fig. A.10) given by[17]

$$c_1^* = q\frac{c_2^* + \sum_{k=2}^{N-1} c_k^*}{p\sum_{k=2}^{N-1} c_k^*} = \frac{3}{2}\frac{q}{p}\,.$$

Hence c_1^* is independent of the initial number of monomers M and is of order q/p. Thus the GFAT depends on M only through the initial condition c_k^*, $k \geq 2$, and is found to be (see numerical results) almost independent of M on several order of magnitude for $N \geq 15$. Finally, note that there is always a (small) probability that the threshold is reached before the metastable period, which is responsible for a bimodal behavior of $T_{\rho,h}^{N,M}$ (see numerical results). For values of d_k^* and a summary of the different cases, see Tables II and I.

Finally, performing a second-order approximation of Eq. (20)-(21), we obtain a system of stochastic differential equation with variance of order $\sqrt{1/M}$ (see details in Annex, section A.3.2), and the GFAT can be computed by

$$MT_{\rho,h}^{N,M} \approx \inf\{t \geq 0 : \tilde{D}_N^M(t) \geq \rho M^{h-1}\}\,,$$

where $(\tilde{D}_k^M)$ denotes the solution of the second order stochastic differential equations. Such approach unfortunately do not provide analytical hints on the behavior of the GFAT, but provide a convenient tool to compute numerically an approximation of the GFAT for very large M, where the exact MC simulations slow down (see numerical results).

2. *"System-size expansion" scaling*

The above asymptotic approximation have been performed assuming that the reaction rates are independent of M. However, the limit $M \to \infty$ may be understood as a system-size expansion, in which case reaction rates must be scaled with the system size according to their respective order. In particular, it is classical[30] that first-order reaction rates are independent of the system size, and second-order reaction rates are inversely proportional to the system size. Thus we are led to use $p_k = \frac{\overline{p}_k}{M}$.

With $\gamma = 0$, the re-scaled variable $D_k^M(t) = C_k(t)/M$ converges now to the BD system given, for all $t \geq 0$, by Eq. (22) and flux definition

$$j_k(t) = \overline{p}_k d_1 d_k(t) - q_k d_{k+1}(t)\,, \quad k \geq 1\,. \tag{27}$$

As before, using the convergence of $(D_k^M(t))$, we obtain the following asymptotic behavior of the GFAT for $h = 1$,

$$\lim_{M\to\infty} T_{\rho,1}^{N,M} = \inf\{t \geq 0 : d_N(t) \geq \rho\}\,.$$

Once again, the latter quantity is deterministic, and may be finite or infinite, according to the respective value of $q_k, \overline{p}_k, N$ and ρ. The GFAT $T_{\rho,h}^{N,M}$ with $h < 1$ behaves asymptotically as the GFAT of the linear CMSBD model (9) with $C_1 \equiv M$ and $p_k = \frac{\overline{p}_k}{M}$. Thus,

$$T_{\rho,h}^{N,M} \approx_{M\to\infty} C(\overline{p}, N)\frac{1}{M^{(1-h)/(N-1)}}\,, \tag{28}$$

where $C(\overline{p}, n)$ is a constant that depends only on N and $\overline{p}(k), k \leq N$. Second-order approximation may also be derived using a central limit theorem for D_k^M (see Annex, section A.3.2).

D. Large M and Large N

In this section, we investigate the behavior of the SBD and its FAT when the size N of the maximal cluster is large, and scales with the total number of particles M. As in section **B**, we will then naturally use the rescale size variable $x = k/N$. We distinguish again two scenario, which yields distinct results. In the first one, the aggregation and fragmentation rates p_k, q_k are taken independent of M. In the second one, the aggregation rates are scaled with the total number of particles, with $p_k = \frac{\overline{p}_k}{M}$. In both cases, a rescaling of the solution is found to be solution of a deterministic continuous size model, namely the Lifshitz-Slyozov model (LS). Indeed, we have detailed in a companion paper[39] how to choose a proper scaling and how to derive the limit equation for that rescaled solution. We show here the consistency of this scaling with the behavior of the GFAT.

We will restrict for simplicity here to the case $N = \sqrt{M}$. In that case, we define the rescaled measure on $\mathbb{R}^+$,

$$\mu^M(t, dx) = \sum_{k\geq 2} \frac{C_k(t/M^\gamma)}{\sqrt{M}}\delta_{k/\sqrt{M}}(dx)\,, \tag{29}$$

and $C_1^M(t) = C_1(t/M^\gamma)/M$, where $\delta_x(\cdot)$ is the Dirac measure at x. The GFAT $T_{\rho,h}^{\sqrt{M},M}$ involves a larger and larger maximal size $\sqrt{M}$, which is rescaled to the macroscopic size $x = 1$ by the definition of the measure μ^M in Eq. (29). We also need to define accordingly macroscopic aggrega-

tion and fragmentation rates, using

$$\left\{\begin{array}{rcl} p^M(x) &=& \displaystyle\sum_{k\geq 2} \overline{p}_k \mathbf{1}_{[k/\sqrt{M},(k+1)/\sqrt{M})}(x)\,, \\ q^M(x) &=& \displaystyle\sum_{k\geq 2} q_k \mathbf{1}_{[k/\sqrt{M},(k+1)/\sqrt{M})}(x)\,, \end{array}\right. \tag{30}$$

where $\mathbf{1}_I(\cdot)$ is the characteristic function that equals 1 in I and 0 outside.

1. $N \to \infty$, No scaling of the aggregation rate

Using $\gamma = 1/2$, and a rescaled nucleation rate[49] $p_1 = \frac{\overline{p}_1}{M}$, we have shown in[39] that the that the measure μ^M satisfies

$$\lim_{M\to\infty} \mu^M(t,dx) = f(t,x)dx\,,$$

where f is a density, solution of the irreversible LS coagulation model (see details in Annex, section 4), given, for all $t \geq 0$, by

$$\left\{\begin{array}{rcl} \displaystyle\frac{\partial}{\partial t} f(t,x) + \frac{\partial}{\partial x}(p(x)c_1(t)f(t,x)) &=& 0\,, \quad \forall x > 0\,, \\ c_1(t) + \displaystyle\int_0^\infty x f(t,x)dx &=& 1\,, \\ \displaystyle\lim_{x\to 0^+}(p(x)f(t,x)) &=& \overline{p}_1 c_1(t)\,, \end{array}\right. \tag{31}$$

with initial condition $c_1(0) = 1$ and $f(0,\cdot) = 0$, and where $p(x)$ is the limit of the macroscopic coagulation rate $p^M(x)$ defined in Eq. (30). Such Eq. (31) is a transport PDE with ingoing characteristics at $x = 0^+$, and is well-defined as a boundary condition at $x = 0$ is given. We refer to the paper[39] for the choice of the boundary condition (that depends on the scaling used in Eq. (29) and the scaling of the reaction rates). The large cluster $C_k(t)$ for $k = \sqrt{M}$ is thus approximated by $f(\sqrt{M}t, 1)$, which yields

$$\sqrt{M} T_{\rho,h}^{\sqrt{M},M} \approx_{M\to\infty} \inf\{t \geq 0 : f(t,1) \geq \rho M^{h-1/2}\}\,. \tag{32}$$

The latter quantity is deterministic, and may be finite or infinite, according to the macroscopic coagulation rate p and the threshold ρ.

2. $N \to \infty$, "System-size expansion" scaling

Finally, if the coagulation rates are rescaled with the system size M, *i.e.* if we choose a rescaled nucleation rate $p_1 = \frac{\overline{p}_1}{M^2}$, and $p_k = \frac{\overline{p}_k}{M}$, $k \geq 2$, then, using $\gamma = -1/2$, we have shown in[39] that the that the measure μ^M statisfies

$$\lim_{M\to\infty} \mu^M(t,dx) = f(t,x)dx\,,$$

where f is a density, solution of the LS coagulation-fragmentation model given, for all $t \geq 0$, by

$$\left\{\begin{array}{rcl} \displaystyle\frac{\partial}{\partial t} f(t,x) + \frac{\partial}{\partial x}\left[(p(x)c_1(t) - q(x))f(t,x)\right] &=& 0\,, \quad \forall x > 0\,, \\ c_1(t) + \displaystyle\int_0^\infty x f(t,x)dx &=& 1\,, \\ \displaystyle\lim_{x\to 0^+}(x^r f(t,x)) &=& c_1(t)\,, \end{array}\right. \tag{33}$$

with initial condition $c_1(0) = 1$ and $f(0,\cdot) = 0$, and where $p(x), q(x)$ are the limit of the macroscopic rate $p^M(x), q^M(x)$ defined in Eq. (30), and $r \in [0,1[$ is determined through the relation $p(x) \approx_{x\to 0} x^r$. Again, such Eq. (33) is well-defined if a boundary condition at $x = 0$ is given when the characteristics are ingoing at $x = 0^+$. We refer to[39] for the choice of the boundary condition (that depends on the scaling used in Eq. (29) and the scaling of the reaction rates). The large cluster $C_k(t)$ for $k = \sqrt{M}$ is now approximated by $f(t/\sqrt{M}, 1)$, so that

$$\frac{1}{\sqrt{M}} T_{\rho,h}^{\sqrt{M},M} \approx_{M\to\infty} \inf\{t \geq 0 : f(t,1) \geq \rho M^{h-1/2}\}\,. \tag{34}$$

The latter quantity is deterministic, and may be finite or infinite, according to the macroscopic rates p, q and ρ.

The results of the last two subsections are illustrated below with the help of numerical simulations. Note that for particular choice of rates p and q, one is able to obtain analytically time-dependent solution of Eq. (31) and Eq. (33) (Annex, section 4.1).

IV. SIMULATIONS AND ANALYSIS

In this section we present results derived from simulations of the SBD model associated to the stochastic Eq. (5)-(6), for various values of its key parameters $\{M, N, p_k, q_k\}$. Specifically, we use an exact stochastic simulation algorithm (kinetic Monte-Carlo, KMC) to calculate first assembly times[42–44]. For each set of $\{M, N, p_k, q_k\}$ we sample 10^3 trajectories (except for few cases where sampling so many trajectories was out of reach in terms of computational time) and follow the time evolution of the cluster populations until the threshold is reached, when the simulation is stopped and the FAT/GFAT recorded. We compare and contrast our numerical results with the theoretical findings of the previous sections. The following is divided in four sections that corresponds to four main results on the FAT and the GFAT. In all Figures from Fig. 3 to Fig. 9 we represent each realization of the FAT (resp. GFAT) in light dot together with its empirical mean in large dot. We superpose on top to it the relevant analytical curves to illustrate the consistency with the theoretical findings.

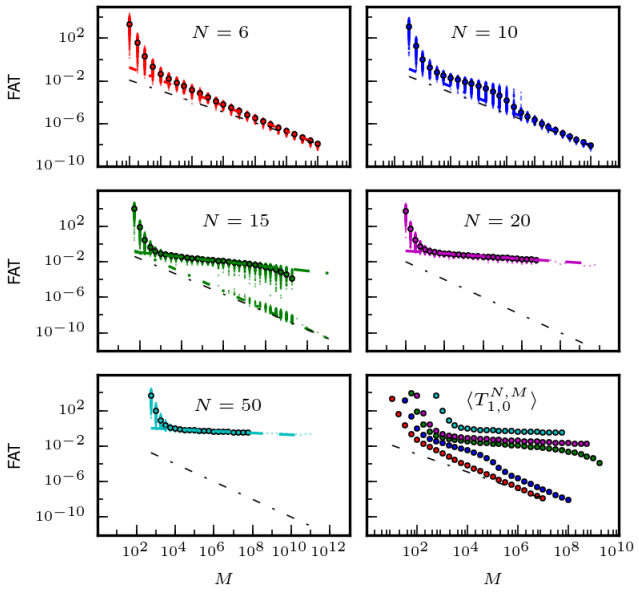

FIG. 3: First Assembly Time $T_{1,0}^{N,M}$ for the original SBD (section III C 1) as a function of the total mass M (in log-log scale) for five different maximal cluster sizes $N \in \{6, 10, 15, 20, 50\}$. Each color light dot is a single realization of the FAT. For each condition, large circles represent the statistical mean over 1000 samples (A few condition are sampled only once, namely for $N = 15, 20, 50$ and large M, for which the mean is not shown). Black dash-dotted lines are straight lines of slope -1, color dash-dotted lines are straight lines of slope $-(1 + 1/(N-1))$ (as in Eq. (14)). And for $N = 15, 20, 50$ we plot additionally dashed lines of slope resp. $-0.26, -0.15$ and -0.10. The last panel in bottom-right represent the 5 mean FAT on the same scale. Kinetic rates are $p_1 = 0.5$, $p_k \equiv 1$, and $q_k \equiv 100$ for all $k \geq 2$.

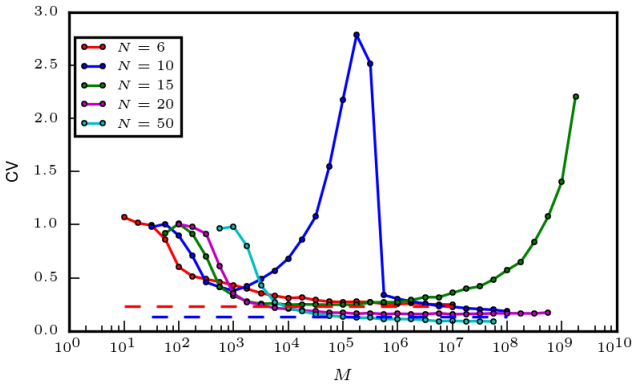

FIG. 4: Coefficient of Variation (CV) for the First Assembly Time $T_{1,0}^{N,M}$ as a function of the total mass M corresponding to the realizations of Fig 3. For $N = 6, 10$ we plot additionally horizontal dashed lines at the value predicted by the Weibull distribution, see Eq. (15).

A. The First Assembly Time can be weakly-dependent on the total number of monomer M, and highly variable even in large population

We begin with the analysis of the FAT as a function of the total number of monomer M, when the maximal cluster size N and the aggregation rates p_k are fixed. For $N = 6, 10, 15$, the prediction of the asymptotic behavior of $T_{1,0}^{N,M}$, the waiting time for a single maximal cluster to be formed, is verified: the mean FAT decrease linearly in log-log scale as M increase, with a slope equal to $-(1 + 1/(N-1))$, as in the linear CMSBD model (Fig. 3, upper panels), see Eq. (14). The coefficient of variation (cv, standard deviation over the mean), that measures the variability of the FAT, is also consistent with a transition from an exponential distribution to a Weibull distribution as M increases: the cv decreases from 1 to the predicted value by Eq. (15) (Fig. 4, and Fig. A.1 in Annex for the CMSBD). Furthermore, one can observe very clearly for $N = 15$ the bimodal behavior predicted for large but intermediate M values (Fig. 3, third panel). For M from 10^6 to 10^{10}, the sampled FAT segregates between two groups separated by several order of magnitude (one group below 10^{-6}, one group around $10^{-2}, 10^{-3}$). The higher values of the sampled FAT corresponds to trajectories that went through the threshold $C_N = 1$ after the metastable period described in paragraph III C 1. For $N = 20$ and $N = 50$, we could simulate in a reasonable computation time (several weeks) only up to $M = 10^{13}$ and $M = 10^{11}$ respectively. Below these values, the metastable states computed in table II predict that the threshold will be mostly reached after the metastable period, which explain the large 'plateau' observed for the FAT up to M^{13} (resp M^{11}): the FAT is nearly independent of M on a broad range of values (Fig. 3, the slope for $N = 15, 20, 50$ are resp. approximatively $-0.26, -0.15, -0.10$). The bimodal behavior observed for the FAT for $N = 10, 15$ can also be visualized on the cv, which results in a large peak of the cv values for intermediate M values (Fig. 4). Trajectories of the number of cluster as a function of time help to visualize the different phases. We illustrate in Annex, Fig. A.7, A.8, stochastic trajectories of the SBD model together with the favorable aggregation limit of the deterministic BD model, in order to clearly identify the metastable period. In Fig. A.9 and A.10, we exhibit two trajectories of the stochastic SBD model that results in two FAT that differ from several log of order of magnitude, due to the metastable period. We also point the accuracy of the approximation by a linear model that has for initial condition the metastable state. Finally, the transition from an exponential distribution to a Weibull distribution as M increases, trough an intermediate bimodal distribution, is also illustrated on Histogramms of the FAT over 10^3 realization in Fig. A.11.

Similar results are obtained for the GFAT $T_{\rho,h}^{N,M}$, where the linear log-dependence with a slope $-(1+(1-h)/(N-1))$ (see Eq. (16) for the CMSBD model) is found to be perfectly satisfied for $N = 3, 5$ and $h = 0.25, 0.5, 0.75$ and

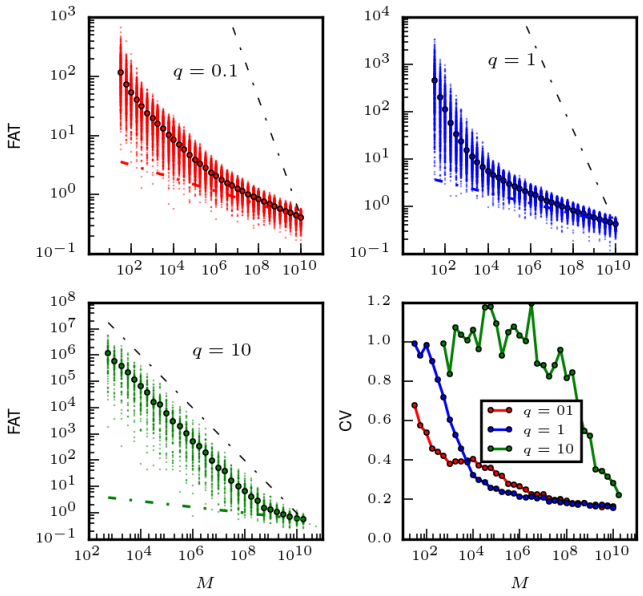

FIG. 5: First Assembly Time $T_{1,0}^{N,M}$ for the rescaled SBD (section III C 2) as a function of the total mass M (in log-log scale) for three different detachment rates $q \in \{0.1, 1, 10\}$, and $N = 10$. Kinetic rates are $p_1 = 0.5$, and $p_k \equiv 1$ and $q_k \equiv q$ for all $k \geq 2$. Each color light dot is a single realization of the FAT. For each condition, large circles represent the statistical mean over 1000 samples. Black dash-dotted lines are straight lines of slope -1, color dash-dotted lines are straight lines of slope $-1/(N-1)$. Finally, the panel in bottom right represent the Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the first three panels (top and bottom left).

$h = 1$ (Annex, Fig. A.5, upper panels). Bimodal behavior and nearly flat log-dependence of the GFAT $T_{\rho,h}^{N,M}$ as a function of M on a broad range of M values is also observed for $N = 10, 20$ (Annex, Fig. A.5, lower panels). The size of the 'bimodal' region is found to be increased with increasing h (and N). For $N = 10, 20$ and $h = 1$, the mean FAT is increasing to ∞ as the deterministic limit given by Eq. (24) is infinite. The cv are non-monotonic with respect to M with a peak corresponding to the bimodal behavior (Annex, Fig. A.6). We show that the GFAT has a lower variability as h increase, and vanish for $h = 1$ and large M, in agreement with the deterministic limit in Eq. (24).

B. The First Assembly Time is non-monotonic with respect to the detachment rate

We verify in Annex, Fig. A.12, A.13 and A.14 the dependence of the FAT on the detachment rate (see also[6]). We confirm that the bimodal behavior is observed for *small* detachment rate, and that the mean FAT (and the cv) is a non-monotonic function of the detachment rate.

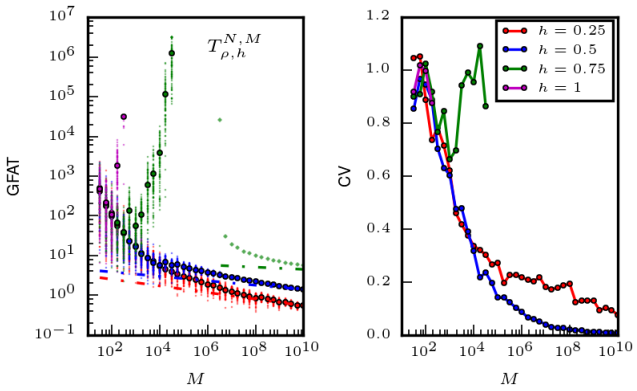

FIG. 6: (Left) Generalized First Assembly Time $T_{\rho,h}^{N,M}$ for the rescaled SBD (section III C 2) as a function of the total mass M (in log-log scale) for $h \in \{0.25, 0.5, 0.75, 1\}$, and $N = 10$. Kinetic rates are $p_1 = 0.5$, $p_k \equiv 1$ and $q_k \equiv 1$ for all $k \geq 2$. Each color light dot is a single realization of the GFAT. For each condition, large circles represent the statistical mean over 1000 samples (A few condition are sampled only once, namely for $h = 0.75$ and large M, for which the mean is not shown). Color dash-dotted lines are straight lines of slope $-(1-h)/(N-1)$. (Right) Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the left panel.

C. The Generalized First Assembly Time may increase with M for the system-size scaling

When the aggregation rates p_k are rescaled with the total number of monomer M (see section III C 2), the FAT to reach a maximal cluster of fixed size N decrease monotonically with M, and asymptotically with a linear log-dependence with a slope $1/(N-1)$ (Fig. 5), as predicted by Eq. (28). The same is valid for the GFAT $T_{\rho,h}^{N,M}$, for $h < 1$, with a slope $(1-h)/(N-1)$ (Fig. 6). However, for $h = 1$, if the threshold ρ is too large, the GFAT is never reached by the deterministic BD model (22)-(27). Thus, for the finite SBD, the GFAT for $h = 1$ increases to ∞ as M increases to ∞. For $h = 0.75$, we also found that the GFAT is *non-monotonic* with respect to the total number of monomers, eventhough it converges to 0 for (very) large number of monomers.

D. Exponentially Large FAT for large maximal cluster size N and phase-transition phenomena

Finally, we verify in Fig. 7 and Fig. 8, that for large maximal size N, of order $\sqrt{M}$, the two scalings show in Eq. (32)-(34) are valid. Specifically, in Fig. 7, we see that for $M > 10^6$, the FAT is nearly deterministic and can then be predicted by the limit model Eq. (31). The same threshold is empirically observed in Fig. 8 for the GFAT as well. However, considering $p(x) = x$ and $q(x) = 1$, in Fig. 9, we show that exponential large devi-

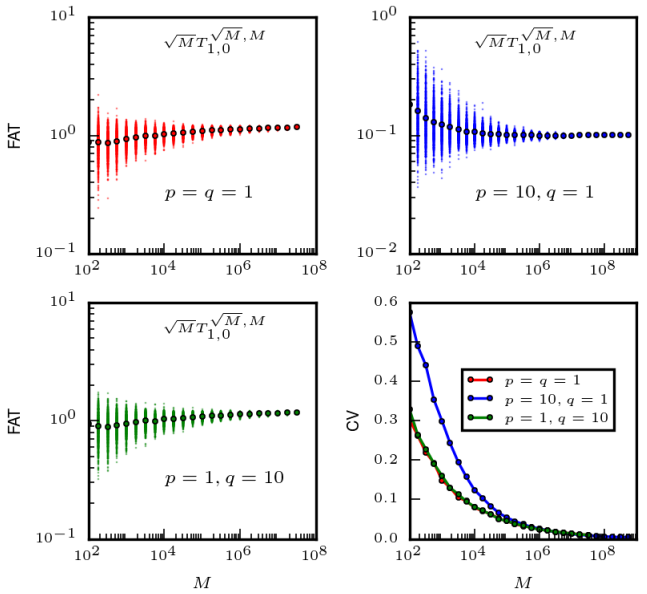

FIG. 7: First Assembly Time $T_{1,0}^{\sqrt{M},M}$ for the original SBD and large maximal cluster size of order $N = \sqrt{M}$ (section III D 1) as a function of the total mass M (in log-log scale) for three different kinetic rates $(p_k, q_k) \in \{(1,1);(10,1);(1,10)\}$. The FAT is multiplied by $\sqrt{M}$ to verify the scaling in Eq. (32). Finally, the panel in bottom right represent the Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the first three panels (top and bottom left).

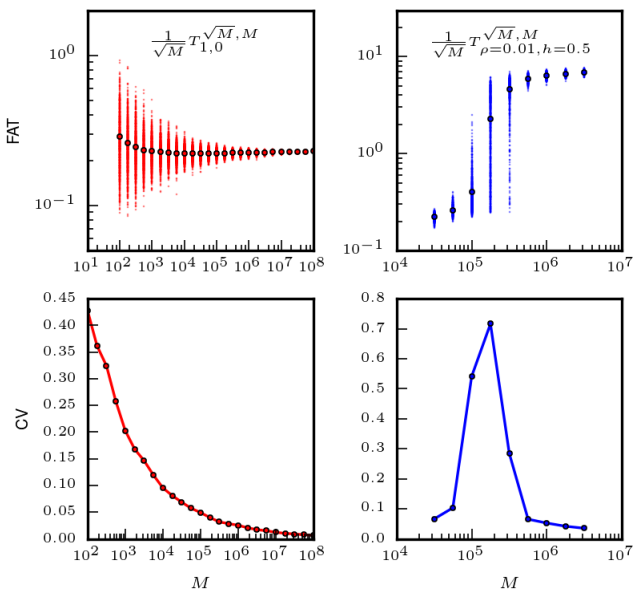

FIG. 8: First Assembly Time $T_{1,0}^{\sqrt{M},M}$ (top left) and Generalized First Assembly Time $T_{\rho,h}^{\sqrt{M},M}$ (top right) for the rescaled SBD and large maximal cluster size of order $N = \sqrt{M}$ (section III D 2) as a function of the total mass M (in log-log scale). Kinetics rates are $p(x) \equiv 5$ and $q(x) = x$. Both the FAT and the GFAT are divided by $\sqrt{M}$ to verify the scaling in Eq. (34). Finally, the panels in bottom left and right represent the Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the upper panels.

ation may occur if the aggregation is not favorable compared to the fragmentation, as in the SCSBD model (17) (Annex, Fig. A.4). Indeed, in such case, the deterministic limit (33) predicts that the FAT is never reached (and equals ∞) as the drift is negative for small (macroscopic) size x. For the finite system, the FAT grows exponentially fast with M, in agreement with Eq. (19). On the right panels in Fig. 9, we show few time-dependent trajectories that are representative of a phase-transition phenomena, with a very abrupt change of phase, occurring at a widely variable time (the cv is near 1). Although the deterministic limit predicts that the aggregation will *not* take place (and the monomer number will not decrease), in the SBD model the aggregation is *always* complete (no monomer at the end), but at larger and larger time as M is increasing.

V. SUMMARY AND CONCLUSIONS

We have studied the problem of determining the First Assembly Time (FAT) of a cluster of a pre-determined size N to form from an initial pool of M independent monomers characterized by size-dependent attachment and detachment rates p_k and q_k, respectively. We have developed a full stochastic approach, based on the stochastic Becker Döring equations (SBD).

We developed two simplified model and were able to find exact results for the FAT statistics for general values of M,N and p_k,q_k. The first simplification is to consider that the number of monomer stays constant over time (linear CMSBD model). The FAT was found to be asymptotically (for large M) a Weibull distribution, and the mean GFAT decrease to 0 as M increase to infinity with a log-linear dependence, with coefficient $1 + (1-h)/(N-1)$, for any $h \in [0,1]$, and any N. The second simplification is to consider that a single cluster can be formed at a time (SCSBD). The FAT was found to be asymptotically an exponential distribution (for large detachment rate q_k), and the mean FAT increase to ∞ as q increase to infinity with a log-linear dependence, with coefficient $N-2$, for any N. We also show that in the case of unfavorable aggregation ($q_k > p_k$), the formation of large cluster takes an exponentially large time as M increases to ∞.

With the analytical results on the simplified model in mind, we analyzed the behavior of the FAT for the full SBD. Using a rescaling strategy, as the total number of monomer M increases to ∞, we found asymptotic expression of the mean FAT and GFAT as a function of a first passage time associated to deterministic models,

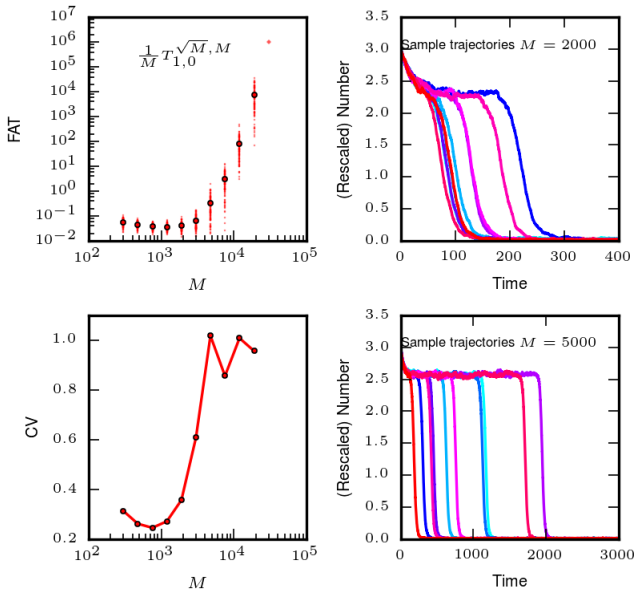

FIG. 9: (Top left) First Assembly Time $T_{1,0}^{\sqrt{M},M}$ for the rescaled SBD and large maximal cluster size of order $N = \sqrt{M}$ (section III D 2) as a function of the total mass M (in log-log scale). Kinetics rates are $p(x) = x$ and $q(x) = 1$. (Bottom Left) Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the upper left panel. (Top and Bottom Right) Time-dependent trajectories of the rescaled number of monomers $c_1(t) = C_1(t)/M$, for $M = 2000$ (top) and $M = 5000$ (down). Each color line represent a single realization with the same initial condition and kinetic parameter.

namely the discrete-size Becker-Döring (BD) model and the continuous-size Lifshitz-Slyozov (LS) model. This has for first implication to be able to find very quickly the order of magnitude of the FAT (resp. GFAT) with the help of a single (fast) numerical simulation of a deterministic model (rather than by extensive numerical simulation of the full SBD model). With the help of a careful time scale analysis on the deterministic BD model, and with extensive numerical simulation, we also pointed out surprising deviations from the mean field deterministic model. First, for sufficiently large maximal cluster size ($N \geq 15$), the mean FAT is found to be very weakly-dependent on the total number of monomers M, and so for several order of magnitude of "intermediate values" of M (from 10^6 to 10^{13} in our simulations). The full distribution of the FAT is bimodal on this parameter region. We explained and gave practical criteria to observe this phenomena by a careful inspection of the metastable state of the favorable aggregation limit for the deterministic BD model. Second, for large maximal cluster size, we confirmed that for unfavorable aggregation kinetic ($q(x) > p(x)$), the mean FAT is exponentially large as M increases to ∞ for the full SBD model. We linked this behavior with phase-transition phenomena, where the number of monomers drastically drops to 0 in a very short time, compared to the FAT. This phase-transition phenomena occurs as a large deviation from the deterministic model, which predicts that the number of monomers stays constant (no aggregation takes place).

This study has generalized previous study on the first passage time on the stochastic Becker-Döring model. Up to our knowledge, this study is the first one to capture the behavior of the FAT and its generalization for arbitrary kinetic rates, and to explore systematically its dependence with respect to the total number of monomers and the size of the maximal cluster. Taking into account size-dependent kinetic rate is important in practice, as monomer binding and unbinding usually depends on the available surface area of the cluster (for the spherical shape, $p_k \sim k^{2/3}$). This study may have several important applications. One of this is the explanation of the nucleation time observed in *in vitro* polymerization assay of misfolded proteins linked to neurodegenerative diseases[11–15]. Typical experiments performed in this field are able to record the nucleation time (defined as the time for which the polymerization starts) for various initial quantity of proteins. Some experiments have described a very weak dependence with respect to this initial quantity, where traditional nucleation theory could not explain this fact. Our stochastic approach points out several new behavior that may explain the observations. Furthermore, we argue that having a model that is able to take into account the observed variability on the nucleation time will be important for parameter inference from experimental data (see also the recent preprint[45]). Indeed, even though the mean FAT may be weakly dependent on the maximal cluster size N (consider the slope of $1 + 1/(N-1)$ for large M), having the observation of the full distribution will facilitate the inference of the maximal cluster size (the shape parameter of the Weibull distribution is $k = N - 1$). Finally, on a more theoretical side, the phase-transition phenomena of the SBD model for unfavorable aggregation and large cluster size seems to be described here for the first time. This gives a possible different definition of the nucleation rate, as an inherent infrequent stochastic process, in contrast to classical nucleation theory. It remains in the future to make a link with studies on gelation phenomena, which happen when a fraction of the mass is concentrated in a giant particle (N is of order of M). Such studies have been performed mostly in general smoluchowski coagulation models[37,46,47].

A number of generalization of this model could be considered and will be relevant to tackle new biophysical problems. One could generalize this study to allow general coagulation-fragmentation between any two clusters[48]. This extension as well as the treatment of heterogeneous nucleation and secondary pathways will be considered in future work.

VI. ACKNOWLEDGEMENTS

This work has been supported by ANR grant MADCOW no. ANR-08-JCJC-0135-01 (France), and Association France-Alzheimer, SM 2014. EH was supported by CAPES/IMPA.

VII. TABLES

TABLE I: **Summary of the First Assembly Time (FAT) and Generalized First Assembly Time (GFAT) findings in this paper. Analytical and numerical results.**

Model	Condition	M (log-log) dep.	Distrib.
CMSBD	$M \to \infty$	$-(1+(1-h)/(N-1))$	Weibull
SCSBD	$q \to \infty$	$-N$	Exponential
SCSBD	$N=\sqrt{M}\to\infty$, $q_k > p_k$	$Me^{\sqrt{M}}$	Expo. Large deviation
SBD	$M \to \infty$	$-(1+(1-h)/(N-1))$	Weibull
SBD	$Md_N^* \ll 1$	~ 0	Bimodal
SBD, $p_k = \overline{p}_k/M$	$M \to \infty$	$-(1-h)/(N-1)$	
SBD	$N=\sqrt{M}\to\infty$	$-1/2$	
SBD, $p_k = \overline{p}_k/M$	$N=\sqrt{M}\to\infty$	$1/2$	
SBD, $p_k = \overline{p}_k/M$	$N=\sqrt{M}\to\infty$, $q_k > p_k$	$Me^{\sqrt{M}}$	Expo. Large deviation

In this table, we sum up the different analytical findings on the FAT, for the full Stochastic Becker-Döring (SBD), Eq. (5)-(6) and its two simplifications with constant monomer (CMSBD), Eq. (9) and single cluster (SCSBD), Eq. (17). The first column denotes which model is considered, with which scaling. The second column provides in which asymptotic the results are valid. The third column gives the slope of the log-log dependence of the GFAT with respect to M (except for the SCSBD and SBD with $N=\sqrt{M}$ where exponential large deviation occurs). The last column gives the full distribution of the FAT (if known). See text for more details.

TABLE II: **Normalized metastable values d_k^* for the deterministic BD model** (22)**-**(26) **in the favorable aggregation case $pM \gg q$.**

size	value	size	value
d_2^*	0.1839	d_{10}^*	9.124010^{-7}
d_3^*	0.1226	d_{11}^*	9.216210^{-8}
d_4^*	0.0460	d_{12}^*	8.448110^{-9}
d_5^*	0.0123	d_{13}^*	7.089410^{-10}
d_6^*	0.0026	d_{14}^*	5.485810^{-11}
d_7^*	4.379510^{-4}	d_{15}^*	3.938510^{-12}
d_8^*	6.386810^{-5}	d_{20}^*	2.873010^{-18}
d_9^*	8.110210^{-6}	d_{50}^*	5.926910^{-64}

In this table, we compute the numerical values of the normalized metastable values d_i^* for the deterministic BD model (22)-(26) with constant kinetic rate $p_k \equiv p$ and $q_k \equiv q$ in the favorable case $pM \gg q$. Such values represent the level that each variable reach during the metastable period after the pure-aggregation period. It is given by the equilibrium value of the irreversible BD model (22)-(23), see Eq. (25). See text in subsection III C 1

[1] R. Becker and W. Döring, Kinetische behandlung der keimbildung in übersättigten dämpfen, *Annalen der Physik* **24** 719-752 (1935).

[2] J. Kuipers, Theory and Simulation of Nucleation, *Utrecht University Repository* **Ph.D.** (2009).

[3] G. M. Whitesides and M. Boncheva, Beyond molecules: self-assembly of mesoscopic and macroscopic components, *Proc. Natl. Acad. Sci. USA* **99** 4769-4774 (2002).

[4] G. M. Whitesides and B. Grzybowski, Self-assembly at all scales, *Science* **295** 2418-2421 (2002).

[5] R. Groß and M. Dorigo, Self-assembly at the macroscopic scale, *Proc. IEEE* **96** 1490-1508 (2008).

[6] R. Yvinec, M. R. D'Orsogna and T. Chou, First passage times in homogeneous nucleation and self-assembly, *J. Chem. Phys.*, **137** 24 (2012)

[7] M. Gibbons, T. Chou, M. R. D'Orsogna, Diffusion-dependent mechanisms of receptor engagement and viral entry, *J. Phys. Chem. B* **114** 15403-15412 (2010).

[8] N. Hoze and D. Holcman, Kinetics of aggregation with a finite number of particles and application to viral capsid assembly, *J. Math. Biol.* **70** 7 1685-1705 (2015).

[9] C. Soto, Unfolding the role of protein misfolding in neurodegenerative diseases, *Nature Rev. Neurosci.* **4** 49-60 (2003).

[10] J. Masela, V.A.A. Jansena and M. A. Nowak, Quantifying the kinetic parameters of prion replication, *Biophys. Chem* **77** 139-152 (1999).

[11] E. T. Powers and D. L. Powers, The kinetics of nucleated polymerizations at high concentrations: amyloid fibril formation near and above the supercritical concentration, *Biophys. J.* **91** 122-132 (2006).

[12] R. Yvinec, Probabilistic modelisation in molecular and cellular biology, *Université Lyon 1* **Ph.D.** tel-00749633 (2012)

[13] E. Hingant, Contributions la modlisation mathmatique et numrique de problmes issus de la biologie - Applications aux Prions et la maladie d'Alzheimer, *Université Lyon 1* **Ph.D.** tel-00763444 (2012)

[14] W.-F. Xue, S. W. Homans and S. E. Radford, Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly, *Proc. Natl. Acad. Sci. USA*, **105** 26 8926–31 (2008)

[15] T. P. J. Knowles et al. An analytical solution to the Kinetics of Breakable filament assembly, *Science*, **1533** 2009 1533–7 (2010)

[16] O. Penrose, The Becker-Döring equations at large times and their connection with the LSW theory of coarsening, *J. Stat. Phys.* **89** 305-320 (1997).

[17] J. A. D. Wattis and J. R. King, Asymptotic solutions of the Becker-Döring equations, *J. Phys. A: Math. Gen.* **31** 7169-7189 (1998).

[18] P. Smereka, Long time behavior of a modified Becker-Döring system, *J. Stat. Phys.* **132** 519-533 (2008).

[19] T. Chou and M. R. D'Orsogna, Coarsening and accelerated equilibration in mass-conserving heterogeneous nucleation, *Phys. Rev. E* **84** 011608 (2011).

[20] F. Schweitzer, L. Schimansky-Geier, W. Ebeling, and H. Ulbricht, A stochastic approach to nucleation in finite systems: theory and computer simulations, *Physica A* **150** 261-279 (1988).

[21] F.P. Kelly, Reversibility and stochastic networks, *Cambridge Mathematical Library* (1979)

[22] A. H. Marcus, Stochastic Coalescence, *Technometrics*, **10** 133-143 (1968).

[23] J. S. Bhatt and I. J. Ford, Kinetics of heterogeneous nucleation for low mean cluster populations, *J. Chem. Phys.* **118** 3166-3176 (2003).

[24] A. A. Lushnikov, Coagulation in Finite Systems, *J. Coll Inter. Scie.* **65** 276-285 (1978).

[25] M. R. D'Orsogna, G. Lakatos, and T. Chou, Stochastic self-assembly of incommensurate clusters, *J. Chem. Phys.* **136** 084110 (2012).

[26] V. Calvez, N. Lenuzza, M. Doumic, J-P Deslys,F. Mouthon and B. Perthame, Prion dynamics with size dependency-strain phenomena, *J. biol. dyn.*, **4** 1751-3766 (2010)

[27] J.M. Ball,J. Carr and O. Penrose, The Becker-Döring Cluster Equations: Basic Properties and Asymptotic Behaviour of Solutions, *Commun. Math. Phys.* , **104** 4 (1986)

[28] P. Laurençot and S. Mischler, From the Becker-Döring to the Lifshitz-Slyozov-Wagner Equations, *J. Stat. Phys.*, **106** 5-6 (2002)

[29] S. Redner, A guide to first passage processes, *Cambridge University Press*, (2001).

[30] N. Van Kampen, Stochastic Processes in Physics and Chemistry, 3rd Edition, *North Holland* (2007)

[31] J. F. C. Kingman, Markov Population Processes, *J. Appl. Prob.* **6** 1-18 (1969).

[32] M. Kreer, Classical BeckerDöring cluster equations: Rigorous results on metastability and longtime behaviour, *Annalen der Physik* , **2** 398–417 (1993)

[33] D. B. Duncan and R. M. Dunwell, Metastability in the Classical, Truncated Becker-Döring Equations, *Proceedings of the Edinburgh Mathematical Society* , **45** 701–716 (dat2002e)

[34] O. Penrose, Nucleation and droplet growth as a stochastic process, in Analysis and Stochastics of Growth Processes and Interface Models, *Oxford University Press*, (2008)

[35] D. T. Gillespie, Transition time statistics in simple bistable chemical systems, *Physica A: Statistical Mechanics and its Applications* , **101** 2 (1980)

[36] D. F. Anderson and T. G. Kurtz, Models of biochemical reaction systems in Stochastic Analysis of Biochemical Systems, *Springer International Publishing* (2015)

[37] I. Jeon, Existence of Gelling Solutions for Coagulation-Fragmentation Equations, *Commun. Math. Phys.* , **567** 541–567 (1998)

[38] N. V. Brilliantov and P. L. Krapivsky, Nonscaling and source-induced scaling behaviour in aggregation model of movable monomers and immovable clusters, *J. Phys. A* , **4789** (1991)

[39] J. Deschamps, E. Hingant and R. Yvinec, Boundary value for a nonlinear transport equation emerging from a stochastic coagulation-fragmentation type model, *arXiv:1412.5025*, (2015)

[40] A. Vasseur, F. Poupaud, J-F Collet and T. Goudon, The Beker-Döring System and its Lifshitz-Slyozov Limit, *SIAM J. Appl. Math.*, **62** 5 (2002)

[41] J-F Collet, Some modelling issues in the theory of fragmentation-coagulation systems, *Commun. Math. Sciences*, **1** 35–54 (2004)

[42] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, A new algorithm for Monte Carlo simulation of Ising spin systems,

J. Comp. Phys. **17** 10-18 (1975).

[43] D.T. Gillespie, Exact Stochastic Simulation of Coupled Chemical Reactions, *J. Phys. Chem.* **81** 2340-2361 (1977).

[44] M. A. Gibson and J. Bruck, Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels, *J. Phys. Chem. A*, **104** 9 1876–1889 (2000)

[45] S. Eugene, W-F. Xue, P. Robert and M. Doumic-Jauffret, Insights into the variability of nucleated amyloid polymerization by a minimalistic model of stochastic protein assembly, *hal-01205549*, (2015)

[46] F. Rezakhanlou, Gelation for MarcusLushnikov process, *Ann. Probab.*, **41** 3B (2013)

[47] N. Fournier and P. Laurençot, Marcus-Lushnikov processes, Smoluchowskis and Florys models, *Stoch. Proc. Appl.*, **119** 1 (2009)

[48] M. R. DOrsogna, Q. Lei and T. Chou, First assembly times and equilibration in stochastic coagulation-fragmentation, *J. Chem. Phys.*, **143** 1 (2015)

[49] The fact that the first aggregation rate need to be rescaled differently from the other aggregation rate comes from the special role played by the monomer in the BD and LS model. For a detailed discussion on the modelling point of view, see[40,41]